

University of Groningen

Divergent or just different

Rozeboom, Henriette

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2014

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Rozeboom, H. (2014). *Divergent or just different: Structural studies on six different enzymes*. [Thesis fully internal (DIV), University of Groningen]. [S.n.].

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Chapter 1

General introduction

ABSTRACT

In this introductory chapter we have chosen to focus on an evolutionary view of protein structure and function. Such a view is of interest as it gives us insight into the inner workings of the evolutionary processes at the molecular level. Understanding how enzymes evolved as biological catalysts to perform the wide variety of reactions found across all kingdoms of life is fundamental to a broad range of biological studies.

The evolutionary process is illustrated with seven examples of enzymes/functional proteins. The three-dimensional structures of these proteins and their evolutionary relationships to other proteins and enzymes are discussed. Furthermore, the technique, X-ray crystallography, by means of which these structures have been determined, is shortly reviewed.

1. BIOLOGICAL EVOLUTION

About 4.6 billion years ago Earth was formed and soon after, 3.6 billion years ago, primordial life began to evolve and the first simple cells appeared (Schopf *et al.*, 2007). How life arose on Earth is still a great mystery. The hypothesis is that life started in a primitive “RNA World” and that later DNA and proteins came into existence (Joyce, 2002). Today, it is generally assumed that the last universal common ancestor of all current living organisms was most likely a single-celled organism that lived ~3.5 billion years ago (Schopf *et al.*, 2007, Woese, 1998). Since then cellular life has split and evolved into the modern life forms belonging to the three ‘domains’ of life we now know as *Archaea*, *Bacteria* and *Eukaryota*.

An important event in the early evolution of life was the emergence of oxygenic photosynthesis, which evolved only once, about 2.4 billion years ago, in a common ancestor of current cyanobacteria that emitted O₂ as a waste product (Schirrmeister *et al.*, 2013). Chloroplasts found in eukaryotes (algae and plants) likely evolved from an endosymbiotic relation with such a cyanobacterial ancestor about 1 billion years ago. Although initially the emitted molecular oxygen (O₂) was captured by dissolved iron and organic matter, later O₂ started to accumulate in the atmosphere. The increasing oxygen concentration forced some obligate anaerobic organisms to adapt, and oxygen-dependent life forms, including multicellular organisms, evolved (Schirrmeister *et al.*, 2013). Nowadays, almost all life on Earth depends ultimately on photosynthetic primary producers, like cyanobacteria and plants (Sleep *et al.*, 2008).

Evolution spans billions of years. Tracing the early evolution of life, by fossil dating can generate clues on the origin of a species. However, identifiable fossil bacteria are not particularly widespread. In addition, it is often very difficult to obtain information about

(extinct) common ancestors. For instance, it is impossible to go far back in time to sequence proteins or DNA from far ancestors that, by definition, no longer exist. In practice, we may bridge a gap of about 100,000-150,000 years as shown by the successful sequencing of the entire genome extracted from the toe bone of a 130,000-year-old Neanderthal woman found in a Siberian cave. According to the genome sequence the *Homo neanderthalensis* is only 0.3% different in its coding sequences from *Homo sapiens* (Prüfer *et al.*, 2014). This research and more to come might shed light on the mystery as to how we evolved and why modern humans are the only surviving human lineage.

2. PROTEIN EVOLUTION

Nowadays, it is generally assumed that all currently existing proteins diverged from a rather small set of proteins (a few hundreds, possibly ~1000), which already existed in the *last universal common ancestor* (Elias *et al.*, 2012). The earliest enzymes were probably generalists, capable of catalyzing particular chemical transformations with broad substrate specificity. During evolution new enzymes emerged by incremental mutations, gene duplications, gene fusion, gene recruitment, gene transfer and post-translational modification (Chothia *et al.*, 2009). Gene duplication provides new genetic material for increasing protein repertoires. The duplicated gene sequence (a paralogue) diverges by mutations, deletions, and insertions to produce modified proteins that may have useful new properties (Chothia *et al.*, 2003). One copy of the gene continues to produce proteins that keep their original, critical function, while the other copy can evolve independently, producing proteins with new and different functions. In general, evolution leads to sequence and structural divergence, which benefits survival at the level of cells and organisms.

An important question concerns the rate of evolution. Is it possible to determine how many years ago two organisms (or proteins) diverged? The *molecular clock* hypothesis proposes that the rate of evolution of a given protein (or, later, DNA sequence) is approximately constant over time and evolutionary lineages (Zuckerkandl *et al.*, 1965, Morgan, 1998). More specifically, the hypothesis proposes that a statistical correlation exists between the time elapsed since the last common ancestor of two contemporary homologous proteins and the number of amino acid differences between their sequences. In practice, this would allow biologists to give a temporal dimension to phylogenetic trees (see below) constructed from molecular data (Morgan, 1998). However, determining the absolute timing of events using molecular clocks has attracted severe skepticism for several reasons. Different proteins appear to evolve at different rates, differences in generation times of organisms may affect the rate, and the size of the population is important as well (in small populations more mutations are neutral)

(Ayala, 1999). All these factors make strict molecular clocks less useful for predictions. Instead, a “relaxed molecular clock” has been proposed that is better able to co-estimate phylogeny and divergence times (Drummond *et al.*, 2006). Another model of evolutionary rate is the “Universal Pacemaker” model (Snir *et al.*, 2012), which is based on the observation that when the rate of evolution changes, the change occurs synchronously in many if not all genes in an evolving genome. These more general models may overcome (partly) the limitations of the molecular clock hypothesis.

Although the rate of mutational change during evolution is, in general, not very precisely defined, the similarities/relationships between organisms, proteins, or DNA sequences can be quantified much more easily. Such evolutionary relationships are often shown in a *phylogenetic tree* or *evolutionary tree*, first sketched by Darwin in his notebook in 1837 (Ragan, 2009). The phylogenetic tree is based upon similarities and differences in the external features and/or genetic characteristics of a certain group of related biological species or gene/protein sequences (taxa). The branches joined together in a node of the tree are implied to have descended from a common ancestor (Drummond *et al.*, 2006).

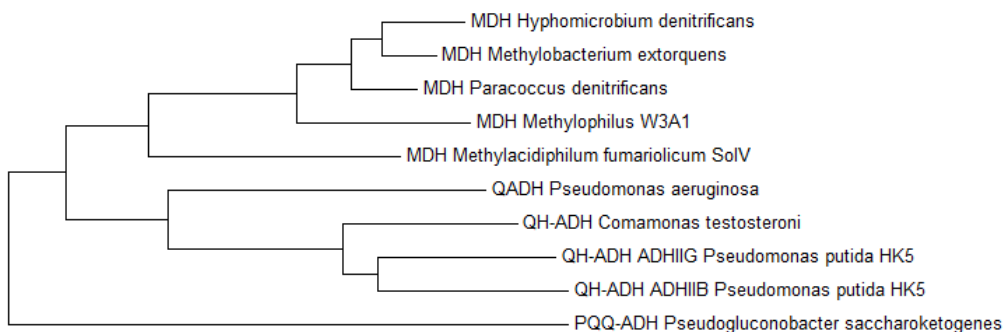


Figure 1. Example of a phylogenetic tree of PQQ-dependent alcohol dehydrogenases obtained with the maximum likelihood method as implemented in MEGA6 (Tamura *et al.*, 2013).

2.1 Evolution and protein folds

The first crystal structures of myoglobin (Kendrew *et al.*, 1958) and, a few years later, of hemoglobin (Perutz *et al.*, 1960) showed that the conformations of the myoglobin and hemoglobin peptide chains are very similar (Cullis *et al.*, 1962, Huber *et al.*, 1971). Although the sequence identity of these two globins is only 24%, they have a striking structural similarity as indicated by their root mean square difference (RMSD) value (see below) of only 1.6 Å. Thus, their 3D structures are much better conserved than their amino acid sequences. The better conservation of 3D structures compared to amino acid

sequences has later been confirmed for many other protein families (Chothia *et al.*, 1986).

Indeed, detecting the evolutionary relationships of proteins from sequence comparisons can be very difficult, especially if their sequences have diverged beyond recognition. In contrast, 3D structures diverge far less rapidly than sequences. Restraints imposed on the tertiary structure of the protein preserve a close-packed hydrophobic interior and conserve the functional properties of the active site (Chothia *et al.*, 1987). Therefore, evolutionary relationship of proteins can often be recognized from their 3D structures, even if the primary structures do not give clear clues (Murzin, 1998).

Similarities in sequence and 3D structure have allowed the classification of proteins. Proteins with significant sequence homology are clustered in a *family*; different families with common structure and function are grouped into protein *superfamilies*. Classification of proteins into superfamilies is detailed in databases such as Structural Classification Of Proteins (SCOP or SCOP2; (Murzin *et al.*, 1995, Andreeva *et al.*, 2014)) and Class Architecture Topology and Homology (CATH; (Sillitoe *et al.*, 2013)). The large amount of structural data coming from structural genomics projects greatly increases the ability to classify protein domains in terms of their evolutionary relationships (Teichmann *et al.*, 2001). The extent of the structural divergence can be quantified by rigid body superposition of the common cores of homologous proteins and calculating the root mean square differences in the positions of their main-chain or C α atoms (Chothia *et al.*, 1986).

2.2 Divergent evolution

Divergent evolution is the evolutionary development of proteins that differ from each other in form and function but that have evolved from the same common ancestor and that usually belong to the same superfamily (Galperin *et al.*, 2012). This type of evolution is most extensively described in the literature and we have chosen to discuss three classic examples of divergent evolution, i.e. the globin family (myoglobins and hemoglobins), lysozymes, and crystallins.

2.2.1 The globin family

As shown above, the structure of hemoglobin revealed a close structural similarity to the structure of myoglobin. Both proteins belong to the very large globin protein family. Globin family members are believed to have evolved from a common ancestor. A detailed analysis of the myoglobin and hemoglobin crystal structures by Lesk and Chothia revealed that mutations of buried residues cause rigid-body movements of the α -helices relative to each other and associated changes in the turn regions, but that the geometry

of the active site is not affected (Lesk *et al.*, 1980). Later, it emerged that nearly all residues in the globin peptide chain can be mutated without compromising the active site (Bashford *et al.*, 1987). This shows the power of evolution: only functional variants are retained. On the other hand, the requirement that function is retained puts severe limitations to the development of new functionalities. It is believed that two whole-genome duplication events (during meiosis) in the globin gene superfamily in the stem lineage of vertebrates were necessary for promoting evolutionary innovation (Storz *et al.*, 2013).

Hemoglobin and myoglobin are good examples of divergent evolution, both probably having arisen from a proto-myoglobin (Aguileta *et al.*, 2006). The single-chain storage protein myoglobin has only one heme group and serves as a reserve for oxygen in muscle and other tissue. The major hemoglobin in adult humans, hemoglobin A (HbA), is a heterotetramer composed of two α -globin and two β -globin polypeptides, each with an associated heme group, which transports oxygen to the cells. The tetrameric organization of hemoglobin tetramer is crucial for its allosteric and oxygen-delivery properties, and the presence of two different globin chains in the tetramer is essential for this function. The α -globin and β -globin chains (42 % sequence identity) are believed to have arisen from the same primordial hemoglobin gene by a gene duplication event, with afterwards each gene following an independent evolutionary history. Other variants of hemoglobin are hemoglobin A2 (HbA2), and fetal hemoglobin F (HbF). The latter hemoglobin consists of two α - and two γ -globin polypeptide chains, which differ at 39 positions from β -globin; HbF has a higher affinity for oxygen because of weaker subunit interface interactions. Moreover, HbF binds the hemoglobin regulator 2,3-biphosphoglycerate less tightly (Frier *et al.*, 1977). These properties ensure delivery of oxygen to the growing fetus. HbA2 is found in trace amounts in the blood and has two α -globins and two δ -globins, which differ at 10 positions from β -globin and which are more resistant to thermal denaturation than HbA. δ -Globins inhibit the polymerization of sickle cell deoxyhemoglobin (HbS) (Sen *et al.*, 2004). γ -Globin and δ -globin evolved from an ancestral β -type globin by further gene duplication. The different isoforms of hemoglobin, resulting from divergent evolution, function according to different tissue requirements for oxygen. Additional heme-containing globins, like cytoglobin and neuroglobin, have been discovered by data mining (Hardison, 2012), but their functional roles are not yet fully clear. Furthermore, plants also contain hemoglobins called leghemoglobins, which, in root nodules, are involved in oxygen transport to symbiotic nitrogen-fixing bacteria or are used as oxygen scavengers/sensors (Hoy *et al.*, 2007).

2.2.2 Lysozyme

Lysozymes are defence enzymes mainly found in egg whites, tears and various secretions of eukaryotic cells. In 1965 hen egg-white lysozyme (HEWL, C (chicken) type) was the very first enzyme of which the atomic structure was solved (Blake *et al.*, 1965). Later, structures were determined of goose type (G type) (Grütter *et al.*, 1983), phage-type (T4) (Matthews *et al.*, 1974) and invertebrate type (I type) (Goto *et al.*, 2007) lysozyme. Matthews & Remington, who solved the 3D structure of T4 phage lysozyme, noticed that its structure was quite different from that of hen egg-white lysozyme (Matthews *et al.*, 1974). However, in 1976 Rossmann & Argos recognized the structural similarity between hen egg-white lysozyme and phage-type lysozyme via rotation and translation superposition of the two structures (Rossmann *et al.*, 1976). This study was the first example of the discovery of an evolutionary relationship of two proteins by comparison of their 3D structures, in the absence of noticeable amino acid sequence homology.

In 1994 Thunnissen *et al.* published the crystal structure of the *E. coli* lytic transglycosylase SLT70 (Thunnissen *et al.*, 1994). SLT70 appeared to have a C-terminal domain with a lysozyme-like fold, which contained the enzyme's active site, where the lytic transglycosylase activity resides. Although the amino acid sequence homology between the four representative lysozyme types (C, G, T4, I) and C-SLT70 is very limited, their three-dimensional structures show striking similarities and indicate their evolutionary relationships (Thunnissen *et al.*, 1995a).

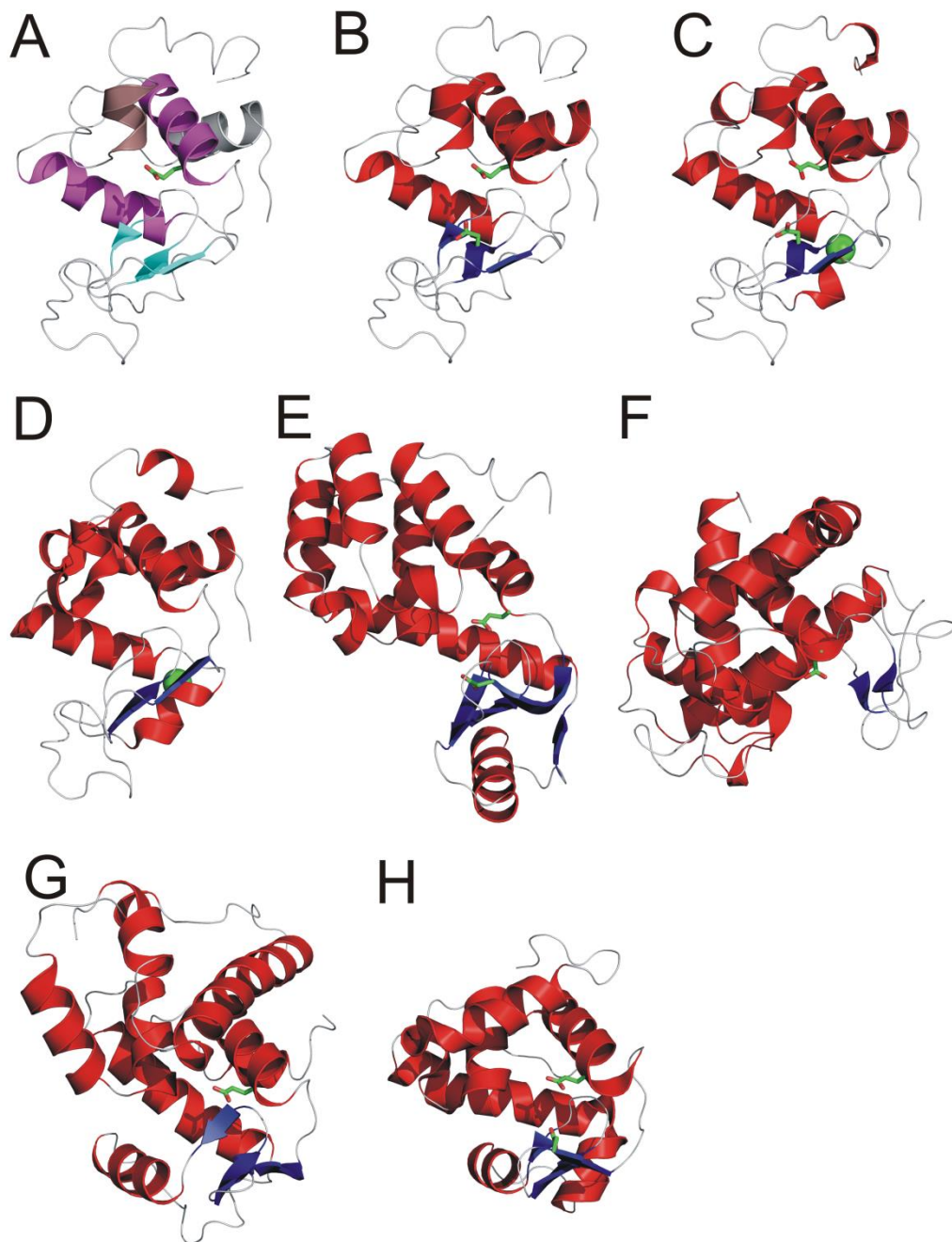
The lysozyme fold is composed of two domains, separated by a deep cleft containing the active site. One domain mainly consists of a β -sheet structure, while the other domain is more helical in nature (Callewaert *et al.*, 2010). The conserved structural core of the lysozyme fold comprises five secondary structure elements that are arranged in a topologically identical order around the substrate-binding cleft. The secondary structure elements comprise a three-stranded β -sheet and two α -helices in the N-terminal lobe and two small α -helical regions in the C-terminal lobe, but with somewhat different orientations in the various proteins (Thunnissen *et al.*, 1995a), similar to what was observed for the globin family (Lesk *et al.*, 1980). Yet, the substrate-binding cleft (where peptidoglycan binds) and part of the active site have been conserved during evolution.

In contrast, the catalytic residues are not fully conserved. Whereas C and I type lysozymes contain two catalytic acidic amino acids in their active site — the acid/base Glu and an Asp that serves as a nucleophile —, in G-type lysozyme and C-SLT70 only the Glu is present. In T4 lysozyme the acid/base catalyst Glu is conserved as well, but this enzyme has also an Asp, which is however not essential for catalysis, and which is at a different position compared to the nucleophilic Asp in C-type lysozyme (Kuroki *et al.*, 1999). These differences clearly suggest that not only the folds of the enzymes have diverged, but also the catalytic mechanisms. Rather than having an enzyme-based

nucleophile, T4 and G-type lysozyme, as well as C-SLT70, make use of the substrate's N-acetyl group to stabilize the reaction intermediate (van Heijenoort, 2011). In SLT70 even the reaction specificity has changed: rather than catalyzing a hydrolysis reaction as the lysozymes do, it catalyzes a $\alpha(1,6)$ -transglycosylation reaction via an intramolecular nucleophilic attack. These examples show that during evolution the substrate-binding sites and the catalytic Glu remained conserved, but that other solutions emerged for the stabilization of the oxocarbenium ion reaction intermediate.

Interestingly, α -lactalbumin, a protein expressed in the lactating mammary gland, has also a lysozyme-like fold, resembling in particular calcium-binding C-type lysozymes found only in a few species of birds and mammals, with which it shares about 35-40% amino acid sequence identity (baboon and human) and four conserved disulfide bridges (Acharya *et al.*, 1989). In α -lactalbumin a Tyr residue blocks the substrate-binding cleft and the catalytic acid/base Glu is not present (Acharya *et al.*, 1989). Gene duplication of a common ancestor and subsequent divergent evolution are thought to be the evolutionary events leading to the coexistence of lysozymes and α -lactalbumins in mammals. Although both proteins have anti-infective activity in common, their functions are quite distinct (Callewaert *et al.*, 2010, Qasba *et al.*, 1997). Whereas lysozymes degrade the bacterial cell wall polymer peptidoglycan, α -lactalbumin is the regulatory subunit of the lactose synthase complex, which is required for the synthesis of lactose in milk. The C-type lysozyme and α -lactalbumin structures represent a clear example of functional divergence preceding gross structural divergence.

Figure 2. Cartoon representation of the crystal structures of (A) the common fold of lysozymes with the two N-terminal lobe helices in magenta, the C-terminal lobe helix in brown, the 3-stranded β -sheet in cyan; the catalytic glutamate and aspartate are depicted as sticks, the green spheres are the calcium ions, (B) *Gallus gallus* (hen egg-white, C type) lysozyme (PDB ID 2LYZ (Blake *et al.*, 1965, Diamond, 1974), (C) *Tachyglossus aculeatus* (Echidna milk) calcium containing lysozyme (PDB ID 1JUG (Guss *et al.*, 1997)), (D) *Papio cynocephalus* (baboon milk) α -lactalbumin (PDB ID 1ALC (Acharya *et al.*, 1989)), (E) Bacteriophage T4 lysozyme (PDB ID 2LZM (Weaver *et al.*, 1987), (F) *Anser anser* (G type) lysozyme (PDB ID 153L (Weaver *et al.*, 1995), (G) *Escherichia coli* soluble lytic transglycosylase C-terminal domain (PDB ID 1SLY (Thunnissen *et al.*, 1995b) and (H) *Tapes japonica* (Japanese cockle, I type) lysozyme (PDB ID 2DQA, (Goto *et al.*, 2007) (for easy viewing the B subunit is not shown).



2.2.3 Crystallins

Crystallins are also very interesting from an evolutionary perspective. They are major lens proteins of the vertebrate eye, and in the eye they are closely packed, and need to be extremely long-lived. They also occur in other tissues, where they have different roles, usually being involved in metabolic processes (Slingsby *et al.*, 2013). This phenomenon, that the same protein fulfills different functions in different cells or organs (also called protein moonlighting), involves the acquisition of a novel secondary function. Usually, this involves a change in gene expression without loss of the primary function, or it involves duplication of the gene (Tsai *et al.*, 2005).

Structurally, α -crystallins are related to small heat-shock proteins (sHSP), which are expressed in many tissues, notably muscle (Slingsby *et al.*, 2013). Remarkably, the same protein is used as α -crystallin in the lens and has a function as chaperone in other tissues (as well as in the lens). Other examples of enzymes moonlighting as a crystallin include α -enolase/ τ -crystallin and argininosuccinate lyase (ASL)/ δ 2-crystallin (Wistow, 1993, Piatigorsky, 2003). This latter case is of some interest, since in ducks two δ -crystallin isoforms exist, δ 1- and δ 2-crystallin, which have 94% amino acid sequence identity. While duck δ 2-crystallin has maintained ASL activity, evolution has rendered duck δ 1-crystallin enzymatically inactive (Tsai *et al.*, 2005). Thus, following the recruitment of ASL as δ -crystallin in an early ancestor of birds and reptiles, adaptive conflict may have provided the selective pressure for *gene duplication*, allowing one gene to make subtle sequence modifications necessary to enhance lens function, while the other maintained enzymatic function (Wistow, 1993). Similarly, the homologous pair S-crystallin/glutathione S-transferase (GST) has also evolved via gene duplications with subsequent divergence (Piatigorsky, 2003). Evidently, structurally stable enzymes have been recruited as lens proteins several times during evolution (Wistow *et al.*, 1988).

2.3 Convergent evolution

Enzymes that belong to different folds or superfamilies and that do not show amino acid sequence homology, but that share the same substrate and reaction specificity and that have strikingly similar active site arrangements are commonly described as the outcome of convergent evolution and are referred to as analogous (Elias *et al.*, 2012). A few examples of convergent evolution are described in the next sections; they are of proteins on which substantial structural information was gathered in the Laboratory of Biophysical Chemistry of the University of Groningen.

2.3.1 Proteases

Convergent evolution of active sites with the same organization of acid-base-nucleophile (Asp-His-Ser) was first identified in the two unrelated *serine proteases* α -chymotrypsin (Matthews *et al.*, 1967) and subtilisin BPN'. The structure of this latter enzyme was independently determined by Kraut *et al.* (Wright *et al.*, 1969) and Drenth *et al.* (Drenth *et al.*, 1971) from crystals obtained from rather different crystallization conditions. Chymotrypsin and subtilisin BPN' have completely different folds, but both feature an active site with Ser, His and Asp as catalytic residues. Their active sites were compared by Matthews *et al.* (Matthews *et al.*, 1977), concluding that in both enzymes the Ser acts as the nucleophile in the reaction, and the Asp-His pair as the binding site for a proton in the transition state. Interestingly, in 1968 Jan Drenth and coworkers had solved the 3D structure of papain (Drenth *et al.*, 1968), a *cysteine protease*, which showed a surprisingly analogous catalytic triad of Cys, His, Asn, but a fold entirely different from that of α -chymotrypsin and subtilisin.

Serine proteases have been classified into 13 clans. The three major clans, based on their structures, are *chymotrypsin/trypsin*-like, *subtilisin*-like and α/β *hydrolase*-like (see also below) (Krem *et al.*, 2001). Chymotrypsin-like enzymes have a β/β structure with a sequence order of the catalytic residues of His-Asp-Ser. Subtilisin-like enzymes have a $\alpha/\beta/\alpha$ structure with the catalytic residues in the order Asp-His-Ser. Finally, α/β hydrolases (e.g. lipases, esterases and *prolyl oligopeptidases*) have a mostly parallel β -structure flanked by α -helices with their catalytic residues in yet another sequence order (Ser-Asp-His). From this short overview it can be concluded that the catalytic triad exists in a variety of forms and evolved independently a number of times (Dodson *et al.*, 1998).

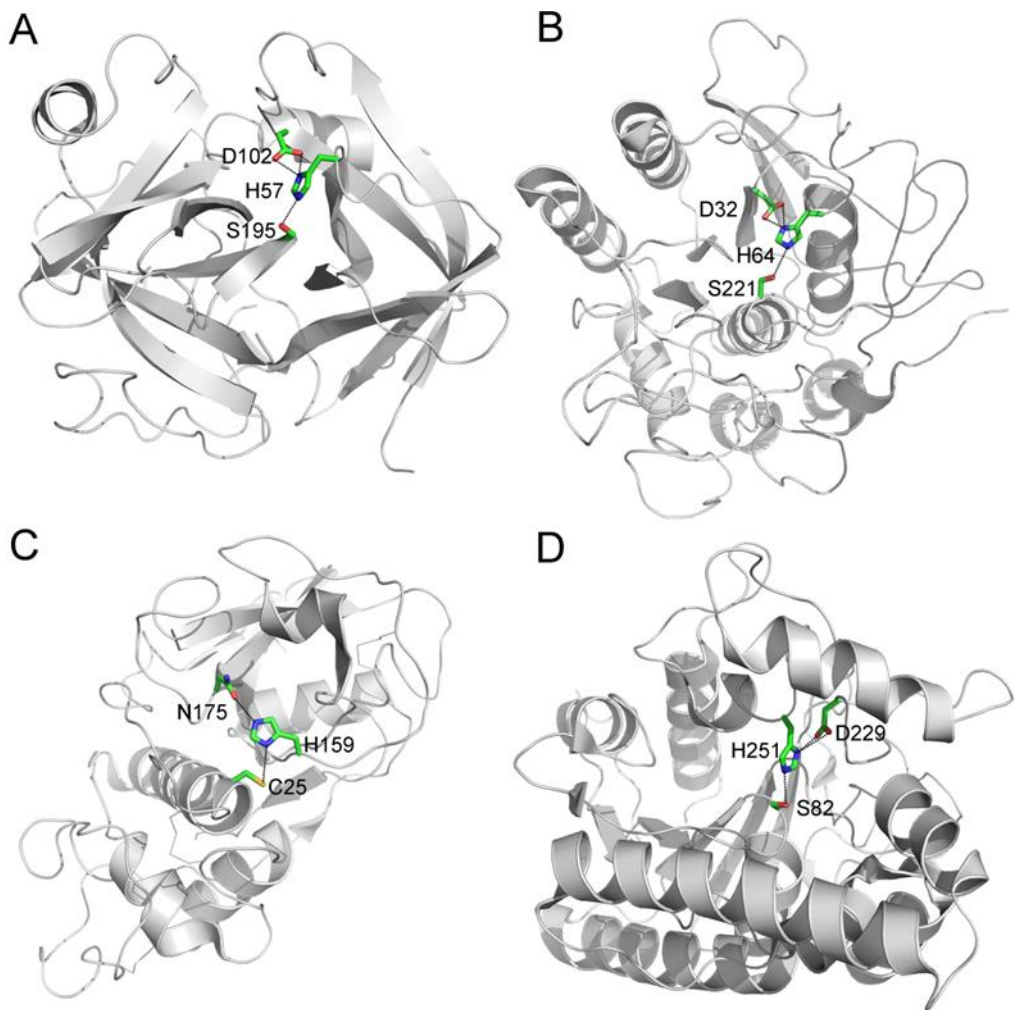


Figure 3. Catalytic triads of three different serine proteases and a cysteine protease. Nitrogen atoms are shown in blue, oxygen atoms are shown in red, sulphur atoms are shown in yellow, H-bonds are indicated by dotted lines. (A) *Bos taurus* chymotrypsinogen (PDB ID 1CHG (Freer et al., 1970), (B) *Bacillus amyloliquefaciens* subtilisin (PDB ID 1SBT (Alden et al., 1971), (C) α/β hydrolase, *Pseudomonas aeruginosa* lipase (PDB ID 1EX9 (Nardini et al., 2000) and (D) cysteine protease, *Carica papaya* papain (PDB ID 1PAD (Drenth et al., 1968).

2.3.2 Hemocyanin - tyrosinase

In contrast to vertebrates, which use hemoglobin as oxygen-carrying protein, some invertebrates deploy hemocyanin to transport oxygen. Hemocyanins are large, blue-colored, multimeric proteins with allosteric oxygen-binding properties similar to hemoglobin (Van Bruggen, 1980). Two different families of hemocyanins exist, *arthropodan* (e.g. arachnids and crustaceans) and *molluscan* (such as squid, cuttlefish, octopus, snails and slugs) hemocyanins. The evolutionary relationship between these hemocyanins has long been a topic of speculation: are they diverged gene products or the result of convergent evolution? Both hemocyanin families have very different amino acid sequences, structures and subunit composition, but there is evidence for a common origin of at least part of their active site (Durstewitz *et al.*, 1997).

Hemocyanins contain a binuclear type-3 copper site that reversibly binds a single oxygen molecule (O_2). This copper site contains two copper ions (Cu-A and Cu-B), each ligated by three histidine residues from different pairs of helices. Because both hemocyanin families contain a similar type-3 binuclear copper site, it is assumed that these proteins both evolved from an ancestral oxygen-binding protein or primitive phenoloxidase (Markl, 2013, van Holde *et al.*, 2001). The binuclear copper site possibly originated from gene duplication and fusion of the Cu-B site. This duplication/fusion event yielded a four-helix bundle, which acquired oxygen-binding ability when O_2 accumulated in the atmosphere and a circulating oxygen-transport protein became crucial to utilize the advantages of aerobic metabolism (van Holde *et al.*, 2001). Although it is likely that gene duplication of a Cu-B binding motif created a Cu-A binding motif in the arthropod hemocyanin evolution, for the molluscan hemocyanin gene fusion of a Cu-B binding motif with an ancestral Cu-A binding site might also have occurred, in view of the quite divergent amino acid sequences of the Cu-A and Cu-B binding motifs in mollusc hemocyanins. Because of this lack of clear sequence conservation in mollusc hemocyanins, while there is in arthropod hemocyanins, this suggests that mollusc and arthropod hemocyanins are the result of convergent evolution.

Both hemocyanin families arose probably independently, with molluscan hemocyanin coming into existence much earlier (about 700-800 million years ago) in evolution than arthropodan hemocyanin (about 550-600 million years ago) (van Holde *et al.*, 2001, Burmester, 2002). The primitive molluscan hemocyanin underwent multiple gene duplications and fusions to yield the current protein, in which 7 to 8 functional units sequentially form a subunit of about ~400 kDa that assembles into a huge decameric or di-decameric cylinder-like quaternary structure (Cuff *et al.*, 1998). For the arthropod hemocyanin precursor the next step in the evolution was the fusion of the α -helical core domain with two other domains, probably involved in cooperative O_2 binding *via* inter-subunit interactions, and the assembly into oligomers of hexamers, with each hexamer

.....

composed of similar or identical subunits with a M_r of ~ 75 kDa (van Holde *et al.*, 2001, Decker *et al.*, 2000). The independent evolution implies that both hemocyanins originated from distinct though similar precursors. The dissimilar evolutionary pathway of the hemocyanins manifests itself mainly at the Cu-A site. In the molluscan hemocyanin structures, one of the copper-coordinating histidine ligands at the Cu-A site is provided from a loop. This is distinctly different from arthropodan hemocyanins, where the corresponding His residue is located on the same α -helix as where the second Cu-A-coordinating histidine resides (Fig. 4) (Cuff *et al.*, 1998). A second distinguishing difference is the presence of a negatively charged residue 4 or 7 residues after the third copper-ligating histidine residue (denoted H_3) of the Cu-A/Cu-B binding motif. In molluscan hemocyanins, a Glu is present at position H_3+7 of the Cu-A binding motif, and an Asp at position H_3+4 of the Cu-B binding motif. In arthropodan hemocyanins, the acidic residue is at position H_3+7 of both the Cu-A and Cu-B binding motifs (Cuff *et al.*, 1998).

Tyrosinases (also called polyphenol oxidases) contain also a type-3 binuclear copper site similar to that of hemocyanins. Since bacterial, fungal, plant, and mammalian (including human) tyrosinases share ~ 25 -30% sequence identity with molluscan hemocyanins, it is generally believed that a primitive molluscan hemocyanin (probably already with some low promiscuous phenoloxidase activity) diverged into present-day molluscan hemocyanin and tyrosinase. Likewise, the primitive arthropodan hemocyanin also diverged into current hemocyanins and prophenoloxidases. Since arthropodan hemocyanins and prophenoloxidases are more similar to each other than molluscan hemocyanins and tyrosinases, the divergence of the latter is more ancient (van Holde *et al.*, 2001). The arthropod-specific prophenoloxidases do not closely resemble their non-arthropod paralogues (Burmester, 2002), indicating a quite different evolutionary history. For instance, the amino acid sequence identity of *Agaricus bisporus* tyrosinase, of molluscan origin, to *Manduca sexta* prophenoloxidase is only 13%. On the basis of sequence similarities two distinct phenoloxidase families can be defined, tyrosinases and prophenoloxidases, respectively from molluscan and arthropodan origin, like the two hemocyanin families.

A related class of proteins, belonging to the arthropod hemocyanin superfamily, is formed by the hexamerins, which are proteins found in insects (arthropods) and which are thought to function as storage proteins, serving as a source of amino acids and energy for protein synthesis during metamorphosis of the larvae (Burmester *et al.*, 1996). Some of them contain a large amount of aromatic amino acids (arylphorins), whereas others have high amounts of methionine (Burmester, 2002). Hexamerins and arthropod hemocyanins share $\sim 30\%$ sequence identity in the core, and likely originate from a common ancestor. However the hexamerins have lost the copper-binding ability because all six copper-ligating histidine residues are mutated (Beintema *et al.*, 1994).

Probably, their oxygen-binding property became obsolete because insects have a tracheal respiratory system using diffusive O₂ transport, and thus have no need for a specific oxygen-transporting protein. Therefore, it has been proposed that hexamerins changed their function to storage after losing the capability of copper-binding (Burmester *et al.*, 1996). However, recently also insect hemocyanins have been described that could be involved in O₂ transport under certain environmental conditions or during some developmental stages (Pick *et al.*, 2009). Likely, the co-existence of copper-less and copper-containing hemocyanin-like proteins may have promoted the emergence of hemocyanin-like storage proteins in insects. Molluscan and arthropod hemocyanins are a clear example of convergent evolution, with two independent gene duplication events, albeit probably starting from the same copper-binding motif. Thereafter, further divergent evolution led to the emergence of tyrosinases, and in the arthropodan family, hexamerins and prophenoloxidasases.

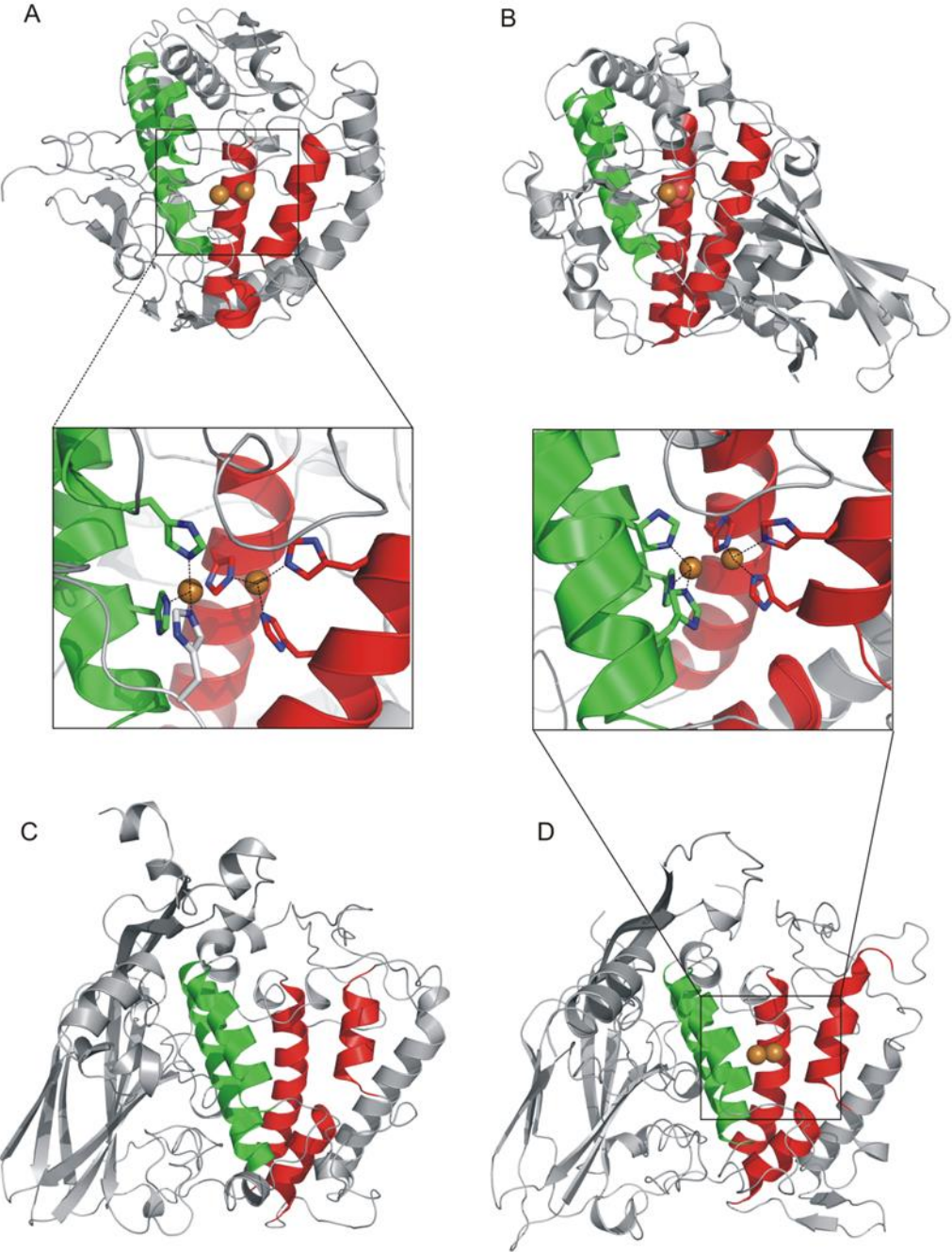


Figure 4. Cartoon representation of the crystal structures of (A) *Agaricus bisporus* tyrosinase (PDB ID 2Y9W (Ismaya *et al.*, 2011)); the His residues, which coordinate the binuclear copper site, are depicted as sticks in the zoomed-in window, (B) *Octopus dofleini* (mollusc) hemocyanin functional unit (PDB ID 1JS8 (Cuff *et al.*, 1998)), (C) *Antheraea pernyi* hexamerin-like arylphorin (PDB ID 3GWJ (Ryu *et al.*, 2009)) and (D) *Panulirus interruptus* (arthropod) hemocyanin (PDB ID 1HC1 (Volbeda *et al.*, 1989). The four α -helices that make up the active site are depicted in green (Cu-A binding motif) and red (Cu-B binding motif). The brown spheres are the Cu-A (left) and Cu-B (right) copper centres. For viewing purposes the lectin-like domain in tyrosinase (A) and the N-terminal domains of (C) and (D) have been omitted.

2.3.3 Other oxygen-transport proteins

A third class of oxygen transport proteins is formed by the hemerythrins. Hemerythrins are responsible for oxygen transport in the marine invertebrate phyla of sipunculids (peanut worms), priapulids (penis worms) and brachiopods (shellfish). Hemerythrin has also a four α -helix bundle core with a pair of iron ions binding O₂. The protein occurs in various multimeric states (monomers, trimers and octamers (Meyer *et al.*, 2010)). The iron ions are bound to the protein through five histidine residues as well as through the carboxylate side chains of a glutamate and an aspartate residue (Holmes *et al.*, 1991). The evolutionary history of hemerythrins is characterized by frequent gene losses (Martín-Durán *et al.*, 2013).

The examples described in sections 2.2.1, 2.3.2, and 2.3.3 show that oxygen-transport proteins have evolved in several independent ways, making use of different metal ions (Fe²⁺ or Cu⁺), and different metal-coordination schemes. From this we conclude that the present-day oxygen-transport proteins are the result of convergent evolution.

2.4 Parallel evolution

Parallel evolution is the appearance of similar functional and structural properties within protein families that evolved in separate lineages from a distant, ancient ancestor. In a remarkable number of cases, parallel evolution has occurred by the repeated acquisition of precisely the same mutations, sometimes in the very same order, which indicates that constraints strongly limit the set of accessible sequences that can produce the selected phenotype (Harms *et al.*, 2013). However, at the DNA and protein sequence level, it is rare to find parallel evolution. Nevertheless, two cases can be noted.

One of these cases is exemplified by the evolution of hemoglobin in birds living in the high altitudes of the Andes and other high-mountain areas. Five lineages of birds can be distinguished that independently evolved hemoglobin variants with a higher oxygen affinity than the hemoglobins of low-altitude birds. These hemoglobin variants evolved in parallel in distantly related lineages. Surprisingly, in all variant hemoglobins the

amino acid substitutions are concentrated in the same few regions of the protein (McCracken *et al.*, 2009).

Parallel evolution has also been observed in haloalcohol dehalogenases. Haloalcohol dehalogenases catalyze the cofactor-independent dehalogenation of vicinal haloalcohols. Crystal structures revealed that they are structurally related to short-chain dehydrogenase/reductase (SDR) enzymes (with up to 30-35% sequence identity), but that they lack the NAD⁺-co-enzyme binding site (Kavanagh *et al.*, 2008). Two families of haloalcohol dehalogenases (HheA/C and HheB) can be discerned that are related to different coenzyme-based SDR subfamilies, indicating that the two dehalogenase families independently originated from two different NAD-binding precursors and diverged via parallel evolution (de Jong *et al.*, 2003).

Finally, as described in section 2.3.2, the development of molluscan hemocyanins and tyrosinases on the one hand and arthropodan hemocyanins, prophenoloxidasases and hexamerins on the other hand can be considered as independent parallel evolution as well.

2.5 Catalytic promiscuity (poly-reactivity)

Catalytic promiscuity is the ability of an enzyme to accelerate another reaction (or reactions) in addition to its biologically relevant one. Promiscuity of enzymes is essential for the evolution of new enzymatic functions. It has been speculated that superfamily ancestors exhibited broad substrate specificity and could possibly catalyze a range of reactions (Khersonsky *et al.*, 2010). Enzymes often possess other, secondary activities. Under selective pressure, the gene for the existing enzyme may be duplicated, followed by the introduction of a small number of mutations. As a result the secondary activity may become a new main activity (Poelarends *et al.*, 2006). The existence of catalytic promiscuity provides the cell with an effective advantage for the survival of the species and its evolution.

A good example of catalytic promiscuity is given by penicillin acylase. Penicillin G acylases (PGAs) are robust industrial catalysts used for the biotransformation of beta-lactams into key intermediates for the production of beta-lactam antibiotics. Their catalytic promiscuity is obvious from the different activities they display, such as transesterification (transformation of an ester into another ester through an interchange of the alkoxy moieties), Markovnikov addition (electrophile addition to a double bond of a nucleophile) or Henry reaction (formation of carbon-carbon bond in a reaction of a nucleophilic nitroalkane with an electrophilic aldehyde or ketone) (Grulich *et al.*, 2013). Catalytic promiscuity should not be confused with broad substrate specificity. For instance, cytochrome P450s and glutathione S-transferases are involved in the detoxification of a wide range of harmful compounds. Being able to detoxify a broad

range of substrates is inherent to their function, and is therefore not to be considered as catalytic promiscuity (Khersonsky *et al.*, 2010, Copley, 2003).

The interesting further examples of catalytic promiscuity reviewed in the next section come from various previous and ongoing research projects of the Laboratory of Biophysical Chemistry of the University of Groningen in collaboration with other research groups.

2.5.1 The α/β hydrolase fold superfamily

The α/β hydrolase fold superfamily of enzymes was proposed on the basis of similarities in 3D structures and conservation of the position of the active site residues with respect to the secondary structure elements (Ollis *et al.*, 1992, Nardini *et al.*, 1999a, Carr *et al.*, 2009). The family contains enzymes that, in principle, catalyze a hydrolysis reaction, although also other activities have been observed such as bromoperoxidase activity (Hecht *et al.*, 1994). The canonical fold of this highly divergent enzyme superfamily consists of a highly twisted, 8–11-stranded, mostly parallel β -sheet. This sheet is flanked on both sides by α -helices. Additional motifs of variable size, structure and position may decorate the core domain (Marchot *et al.*, 2012). These enzymes have diverged from a common ancestor and preserve the arrangement of the catalytic residues, but not the substrate-binding site. They all have a catalytic triad, the elements of which are borne on loops, which are the best-conserved structural features of the fold (Ollis *et al.*, 1992, Carr *et al.*, 2009).

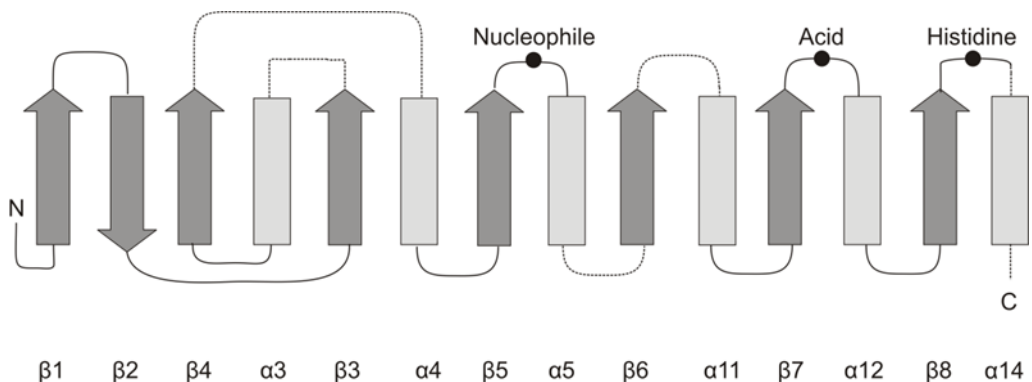


Figure 5. The canonical α/β hydrolase fold. α -Helices and β -strands are represented by rectangles and arrows, respectively. Locations, where insertions in the canonical fold may occur, are indicated by dashed lines.

Enzymes belonging to this superfamily include serine carboxypeptidases, peroxidases, esterases, lipases (Nardini *et al.*, 2000, Lang *et al.*, 1998, van Pouderoyen *et al.*, 2001,

Tiesinga *et al.*, 2007, Rozeboom *et al.*, 2014), cutinases, thioesterases, diene lactone hydrolases, epoxide hydrolases (Nardini *et al.*, 1999b) and haloalkane dehalogenases (Franken *et al.*, 1991, Verschuere *et al.*, 1993, Ridder *et al.*, 1999).

Especially the lipases of the superfamily have been widely applied in biotechnology because of their broad substrate-specificity, and regio- and stereospecificity. (Jaeger *et al.*, 2002). Lipases have also been shown to catalyze other (promiscuous) reactions such as formation of carbon-carbon, carbon heteroatom, and heteroatom-heteroatom bonds through non-hydrolytic reactions, such as aldol condensations and Michael-type additions, and amide hydrolysis by cleavage of a different bond (C-N versus C-O) (Kapoor *et al.*, 2012). Because of their broad repertoire of catalyzed reactions several industrial processes based on lipase catalysis have been established (Jaeger *et al.*, 2002).

2.5.2 Tautomerases

The tautomerase/MIF superfamily members catalyze a keto-enol *tautomerisation* reaction, promoted by an amino-terminal proline; they all exhibit a two-layered β - α - β -fold structure. The superfamily can be divided into five families, each named after a representative reference enzyme, viz. macrophage migration inhibitory factor (MIF), malonate semialdehyde decarboxylase (MSAD), 4-oxalocrotonate tautomerase (4-OT) (including *trans*-3-chloroacrylic acid dehalogenase, CaaD), *cis*-3-chloroacrylic acid dehalogenase (*cis*-CaaD), and 5-carboxymethyl-2-hydroxymuconate isomerase (CHMI) (Baas *et al.*, 2013). Although members of the five families have little overall sequence similarity, the catalytic amino-terminal proline residue is absolutely conserved in all sequences (Poelarends *et al.*, 2008a). Most family members are homotrimers, only CaaD is a heterohexamer with 3 α - and 3 β -subunits, and 4-OT is a hexamer (Poelarends *et al.*, 2008a).

Several members of the superfamily display promiscuous activities. For instance, *trans*-CaaD and *cis*-CaaD primarily catalyze the hydrolytic dehalogenation of *trans*-3-chloroacrylate and *cis*-3-chloroacrylate, respectively, to yield malonate semialdehyde. Both enzymes have catalytic promiscuity in that they also catalyze the tautomerization of phenylpyruvate to phenylpyruvate and the hydration of 2-oxopent-3-ynoate (2-OP) to acetopyruvate (Baas *et al.*, 2013). Both enzymes seem to have evolved within a few decades and appear to be products of rapid evolution (Baas *et al.*, 2013). This suggests a very fast ticking molecular clock under the evolutionary pressure of the xenobiotic nematocide 1,3-dichloropropene.

Another example of a promiscuous enzyme in the superfamily is Cg10062 from *Corynebacterium glutamicum*. Cg10062 is a *cis*-CaaD homologue, which catalyzes the hydrolytic dehalogenation of both *cis*- and *trans*-3-chloroacrylate. In addition, the enzyme has hydratase activity, but its major activity is the dehalogenation of *cis*-3-

chloroacrylate (Baas *et al.*, 2013); the *trans*-CaaD and hydratase activities can be considered as promiscuous activities. Because of its 34% sequence identity to *cis*-CaaD, the enzyme has been suggested to be an intermediate in *Corynebacterium glutamicum* in the evolution to a fully functional *cis*-isomer-specific dehalogenase such as *cis*-CaaD (Baas *et al.*, 2013, Poelarends *et al.*, 2008b). The promiscuous activities provide evidence for *divergent* evolution from a so far unknown common ancestor.

An interesting and perhaps unexpected member of the tautomerase superfamily is the macrophage migration inhibitory factor (MIF). The mammalian MIF factor plays an important role in inflammatory and immune-mediated diseases and triggers an acute immune response. Mammalian MIF is also known as a very efficient phenylpyruvate tautomerase and has promiscuous *trans*-3-CaaD activity (Baas *et al.*, 2013). The MIF homologue from the marine cyanobacterium *Prochlorococcus marinus* has also tautomerase activity and shares mechanistic and structural similarities with mammalian MIFs (Wasiel *et al.*, 2010). As mammalian MIF was presumably recruited long ago to serve a non-enzymatic function, it is quite remarkable that it still exhibits promiscuous enzymatic activities typical of the other superfamily members (Poelarends *et al.*, 2008a). This suggests that the cyanobacterial MIF and mammalian MIFs are related by *divergent* evolution. The question arises whether a MIF-like protein from a free-living bacterium also possesses immunostimulatory features similar to those of mammalian MIFs (Wasiel *et al.*, 2010), but we are not aware of any experiments aiming to shed light on this question.

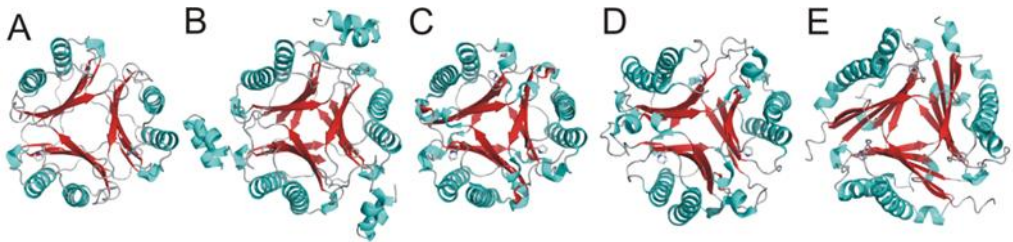


Figure 6. Ribbon diagrams of crystal structures of members of the tautomerase family. (A) *Prochlorococcus marinus* MIF (PDB ID: 2XCZ (Wasiel *et al.*, 2010), (B) *Corynebacterium glutamicum* Cg10062 (PDB ID: 3N4G (unpublished), (C) *Pseudomonas pavonaceae* CaaD (4-OT) (PDB ID: 1S0Y, (de Jong *et al.*, 2004), (D) *Pseudomonas pavonaceae* MSAD (PDB ID: 2AAG (Almud *et al.*, 2005), (E) *E. coli* CHMI (PDB ID: 1OTG (Subramanya *et al.*, 1996).

2.6 Remaining questions

Experimental resurrection of ancestors of universal proteins that have been maintained in nearly all organisms throughout the history of life, together with the sensitivity of single-molecule techniques, can be a powerful tool for understanding the origin and

.....

evolution of life on Earth. *Paleoenzymology* is the study of prehistoric life through the properties of enzymes. Important pathways such as energy production, sugar degradation, cofactor biosynthesis and amino acid processing are highly conserved from bacteria to humans and were probably already present in the last universal common ancestor.

Resurrection of extinct proteins is aimed to obtain information on the effects of historical mutations on functional diversification. For example, in a thioredoxin (Trx) resurrection study sequences predicted by phylogenetic analysis were used for reconstruction of probable variants. It appeared that ancient Trxs use chemical mechanisms of reduction similar to those of modern enzymes but have adapted to their environment from ancient to modern Earth by adapting their activity to the temperature and pH conditions of present-day Earth (Perez-Jimenez *et al.*, 2011). Additional reconstruction studies with other enzymes may provide additional insights on the evolution of ancient phenotypes and metabolisms.

Experimental determination of protein function lags far behind the rate of sequence and structure determination. Nature's strategies for evolving catalytic functions can be deciphered from the information contained in expanding protein sequence databases. However, the functions of many proteins in the protein sequence and structure databases are either uncertain (too divergent to assign function based on homology) or unknown (no homologs), thereby limiting the utility of the databases. Furthermore, these databases are probably far from complete, given the current rate of discovery of new sequences from the metagenome (Kahn, 2011). Nevertheless, the enormous enzyme diversity leaves a huge amount of effort to be done on enzyme functional characterization, to allow further advances in medicine, chemistry, synthetic biology, and industry.

The structural information presented in this thesis is a pertinent contribution of how enzymes can evolve. The 3D structures discussed in the thesis provide insights how altered active sites in α -1,4 glucan lyase, endo-xylogalacturonan hydrolase, and pyrroloquinoline-quinone-dependent alcohol dehydrogenase change the product and substrate specificity compared with homologous enzymes. The structures of the esterases reveal how enantioselectivity can arise through divergent evolution. Furthermore, the structure of mannuronan epimerase shows that its catalytic machinery has evolved through convergent evolution.

3 SCOPE OF THIS THESIS

3.1 X-ray crystallography

For understanding the catalytic mechanisms of enzymes, high-resolution structures as determined by X-ray crystallography are a prerequisite. Nowadays, X-ray diffraction techniques have passed the state of pioneering techniques. Although still a highly technical process, it is no longer abstruse and time-consuming (Morin *et al.*, 2013). Yet, there is still nothing routine about determining the complete three-dimensional structure of a protein (Alberts *et al.*, 2002).

Early X-ray crystallographers referred to attaining the 3D-coordinates of a protein as “*solving*” the structure, reflecting their perception of the process as a kind of intricate scientific puzzle of many parts. Today, the preferred terminology is “*structure determination*”, denoting the more prescribed and rigorous methods typically employed (Morin *et al.*, 2013). By the end of 2013 there were almost 100,000 structures deposited in the PDB (Protein Data Bank) including about 85,000 determined by X-ray crystallography. The number of structures deposited each year in the PDB is growing exponentially since the year 2000, when only 10,696 structures had been deposited.

X-ray crystallography is utilized as a tool to study macromolecular structures and to aid in elucidating the mechanism of action of enzymes. However, the structure-function analysis includes more than just crystallization, X-ray crystallography and structure analysis. Proteins have to be produced (cloning, bacterial culture growth, overexpression, mutagenesis). The produced proteins have to be purified by means of chromatography (affinity, cation/anion exchange chromatography, gel filtration, etc.). Activity tests and enzyme kinetics and inhibition experiments are often done, and conformation analysis is carried out using, for instance, dynamic light scattering and SAXS (small-angle X-ray scattering). All these experiments serve to obtain a protein preparation of the highest quality.

X-ray crystallography requires that the protein be subjected to conditions that cause the molecules to aggregate into a large, perfectly ordered crystalline array, a protein crystal. Each protein behaves quite differently in this respect, and protein crystals can be generated only through exhaustive trial-and-error methods that often may take many years to succeed, if they succeed at all (Alberts *et al.*, 2002, Van der Klei *et al.*, 1989, Boys *et al.*, 1989). The first enzyme molecule to be crystallized in 1926 was jack bean urease by the American biochemist J. B. Sumner for which he was awarded the Nobel Prize in chemistry in 1946. It took another 83 years before its crystal structure was finally reported (Balasubramanian *et al.*, 2010). Nevertheless, protein crystallization successes

.....

have multiplied rapidly in recent years owing to the advent of practical, easy-to-use screening kits and the application of laboratory robotics (McPherson *et al.*, 2014).

Nowadays, protein crystals are scooped up by a loop, which is made of nylon or plastic and attached to a solid rod. Crystals are pre-soaked in a cryoprotectant, and then flash-cooled with liquid-nitrogen to prevent radiation damage. As X-ray sources the synchrotrons in Grenoble (ESRF) and Hamburg (EMBL outstation at DESY) are used, complemented with the in-house Cu-K α radiation source, equipped with a MarDTB Goniostat and DIP2030H image plate detector. The major challenges to determine the phase of each reflection in the diffraction data sets of the proteins discussed in this thesis have been overcome either by multiple isomorphous replacement (MIR, soaking crystals in solutions of heavy atoms (see chapter 2)) or by Molecular Replacement (MR, making use of previously determined structures of homologous proteins as phasing models (see chapters, 3, 4, 5 and 6)). Model building and phase refinement is done with freely available programs and as a final step the atomic coordinates of the final models and structure factors are deposited in the PDB.

3.2 Enzymes

The enzymes discussed in this thesis are a heterogeneous group of enzymes, although each protein is of interest by itself because of its unique way of substrate conversion. The six different proteins, each belonging to a different class of enzymes, specifically belong to the classes of oxidoreductases (E.C.1 ~9.000 structures in the PDB), hydrolases (E.C.3 ~20.000), lyases (E.C.4 ~ 4.000) and isomerases (E.C.5 ~2.500). They are grouped in folds, the α/β hydrolase fold (**Chapter 6**) and the TIM (β/α) $_8$ -barrel (**Chapter 3**). Other folds discussed are the 8-bladed β -propellers (**Chapter 4**) and the right-handed β -helix (**Chapters 2 and 5**).

3.2.1 Evolution of reaction specificity (different products)

3.2.1.1 Mannuronan C-5-epimerase (*AlgE4*)

In *Azotobacter vinelandii* a family of 7 secreted and calcium-dependent mannuronan C-5 epimerases (AlgE1-7) has been identified (Aachmann *et al.*, 2006). Alginate is a family of linear copolymers of (1 \rightarrow 4)-linked β -D-mannuronic acid (M) and its C-5 epimer α -L-guluronic acid (G). The polymer is first produced as polymannuronic acid and the guluronic acid residues are then introduced at the polymer level as blocks by mannuronan C-5-epimerases. This reaction is catalyzed by mannuronan C-5-epimerase, which abstracts the proton at C-5. The reaction is completed by re-donation of a proton to C5 from the opposite face of the sugar ring.

.....

The smallest of the AlgE proteins, AlgeE4, comprises one A- followed by one R-module and strictly forms alternating MG sequences (MG-blocks). The A-modules alone are catalytically active, but their reaction rates are significantly increased when covalently bound to at least one R-module (Ertesvåg *et al.*, 1999). The structure of the A-module, a right-handed parallel β -helix structure was elucidated by X-ray crystallography (**Chapter 2**). The catalytic machinery displays a stunning resemblance, presumably through *convergent* evolution, to the structurally unrelated alginate lyase A1-III from family PL5 that contains an $(\alpha/\alpha)_6$ barrel and alginate lyase ALY-1 from family PL7 containing a β -jelly roll.

3.2.1.2. α -1,4 Glucan lyase (GLase)

The crystal structure of GLase is described in **chapter 3** in the native form and in complex with the inhibitors 1-deoxynojirimycin, acarbose, 5-fluoro- β -L-idopyranosyl fluoride (5F β IdoF) and 5-fluoro- α -D-glucopyranosyl fluoride (5F α GlcF). The unique lytic enzyme α -1,4-glucan lyase (EC 4.2.2.13) has been assigned to glycoside hydrolase family GH31, which consists of mainly hydrolases (EC 3.2.1.x). The enzymes contain a $(\beta/\alpha)_8$ -barrel as core domain with the active site at the C-terminal ends of the barrel strands. The $(\beta/\alpha)_8$ -barrel, which was found first in triosephosphate isomerase and therefore is also known as TIM barrel, is the most common enzyme fold. $(\beta/\alpha)_8$ -Barrels have high structural homology but their overall sequence identities can be insignificant. Structural homology suggests a common evolutionary origin for the four major classes of GH31, which, however, have *diverged* to meet the substrate specificity as desired for their environment. GLase represents an example of a way in which enzymatic mechanisms can evolve through subtle changes in active site constitution.

3.2.2 Evolution of substrate specificity

3.2.2.1 Pyrroloquinoline quinone dependent alcohol dehydrogenase (PQQ-ADH)

Pyrroloquinoline quinone dependent alcohol dehydrogenase (PQQ-ADH) belongs to the E.C. 1.1.5 family of oxidoreductases. Its 3D structure elucidation and characterization are described in **chapter 4**. PQQ-ADH belongs to the PQQ family of β -propeller-fold proteins. β -Propeller proteins adopt folds comprising 4 to 10 repeats of a 4-stranded β -meander called a 'blade'. They are thought to have largely arisen by the independent amplification of one ancestral blade and not from an ancestral propeller. The differentiation of the first proto-propellers yielded a population of new β -meanders, which could serve as starting points for further amplification. This evolutionary pathway of amplification and diversification is an ongoing process (Chaudhuri *et al.*, 2008).

The phylogenetic relationship of the PQQ-domains of ten structurally characterized PQQ-dependent ADHs with eight 4-stranded β -meanders is shown in Fig. 1. Type II QH-ADHs are structurally closely related to QADHs, from which we infer that they are also evolutionary related. Somewhat more distant from these types are type I Methanol Dehydrogenases. Our results show that PQQ-ADH is the most distant to all mentioned types and could be regarded as an outgroup in this tree. This suggests that PQQ-ADH branched from the universal ancestor before the other enzymes branched from each other. Yet, the number of samples is limited in this particular example and the tree may be altered when new 3D structures become available.

3.2.2.2 *Endo-Xylogalacturonan hydrolase (XghA) from Aspergillus tubingensis*

XghA belongs to E.C. 3.2.1.-, (no subfamily assigned) and is a member of Glycoside Hydrolase family GH28, which contains polygalacturonases and rhamnogalacturonases. Family GH28 is a set of structurally related enzymes with interesting functional diversity that hydrolyze glycosidic bonds in pectin and exhibit the parallel β -helix topology. The β -helix fold shows extensive stacking of side chains of similar residues in its interior to give aliphatic, aromatic and polar stacks such as the asparagine ladder. This suggests that coils can be added or removed by duplication or deletion of the DNA corresponding to one or more coils and explains how homologous proteins can have different numbers of coils (Jenkins *et al.*, 2001). In most β -helix proteins a surface groove is observed that forms the center of the active site, but the active site is also shaped by additional loops. The structure of Mannuronan C-5 epimerase also contains a β -helix (chapter 2) with a surface groove but the catalytic machinery is completely different.

Aspergillus tubingensis and *Aspergillus niger* are closely related fungi, which both belong to the subgenus *Circumdati*, section *Nigri* (Abarca *et al.*, 2004). In *A. niger*, 20 putative GH28 homologs are found (Sprockett *et al.*, 2011). The evolution of the GH28 gene family within the fungal kingdom is consistent with the birth-and-death model. In the birth-and-death model new genes are created by repeated gene duplication and some duplicate genes are maintained in the genome for a long time but others are deleted or become nonfunctional by deleterious mutations. The initial appearance of GH28 enzymes predates the evolution of fungi, which places the origin of this gene family at more than 1.5 billion years ago. The earliest fungi already possessed much of the functional diversity seen today in GH28, including exo-polygalacturonan-, endo-polygalacturonan-, and endo-rhamnogalacturonan-acting enzymes (Sprockett *et al.*, 2011).

In **chapter 5** the structure of XghA is described. XghA has probably evolved through gene duplication of an ancestor of GH28 enzymes followed by *divergent* evolution. Its

.....

substrate-binding cleft can accommodate xylosylated substrates and is much wider than in other polygalacturonases that cannot process xylosylated polygalacturonan.

3.2.3 Evolution of enantioselectivity

3.2.3.1 Esterases (NP and CesB)

Examples of enzymes of which the enantioselectivity has successfully been improved or inverted by rational design are rare. The effects of certain mutations to alter enzyme enantioselectivity are more difficult to predict than mutations focused on altering other catalytic properties. However, directed evolution (evolutionary biotechnology or molecular evolution) combined with rational design can be used to improve the enantioselectivity (Bornscheuer *et al.*, 2001, Boettcher *et al.*, 2010).

The carboxylesterases naproxen esterase (NP) and CesB perform enantioselective biochemical transformations. NP was characterized as a very efficient enantioselective biocatalyst for the kinetic resolution of esters of the non-steroidal anti-inflammatory drug naproxen, while CesB shows promise for application in the enantioselective production of the chiral alcohol 1,2-O-isopropylidene-*sn*-glycerol.

Although the crystallization of NP from *Bacillus subtilis* Thai I-8 was already reported in 1993 (van der Laan *et al.*, 1993), it took 20 years to determine its structure, as described in **chapter 6**. In addition, in this chapter the 3D structure of its homolog CesB from *B. subtilis* 168 is described. Carboxylesterases NP and CesB belong to class EC 3.1.1.1 (hydrolases) and contain a α/β hydrolase fold with an α -helical cap domain. Many subclasses of the esterases have been described and many structures have determined so far. Both of the determined esterases show structural flexibility necessary to perform their activity. This makes the outcome of just rational design of mutants that should increase enantioselectivity difficult to predict.

Chapter 7 summarizes the main conclusions of the research described in this thesis. Together with the findings from biochemical and functional experiments, the lessons learned from the structures of glucansucrases and fructansucrases provide directions for future research on these intriguing enzymes.

